Machine Learning with Graphs: Network Science 2/2 - Network Models

Arlei Silva

Spring 2022

The Erdos-Renyi Model

In the last lecture, we have learned about different structural properties of real graphs, such as the degree distribution and clustering coefficient. Here, we will focus on network models that are able to reproduce the properties of real networks. There are several reasons why such models are useful, for instance:

- They enable the generation of large synthetic datasets;
- They support mathematical reasoning about properties of a family of graphs;
- They might provide evidence for the link formation process of the data.

The simplest class of network models are *random graph models*. The most popular among these models is the *Erdos-Renyi (ER) model* $G(n, p)$, where n is the number of vertices and edges are formed with uniform probability p. Using this simple definition, we can infer the following properties of ER graphs:

Probability of any network G with n vertices and m edges:

$$P(G) = p^m (1-p)^{\binom{n}{2}-m}$$

Expected number of edges:

$$\mathbb{E}[m] = \binom{n}{2} p$$

Expected degree:

$$\mathbb{E}[deg(v)] = (n-1)p$$

Degree distribution:

$$P[deg(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{1}$$

which is the *Binomial distribution*.

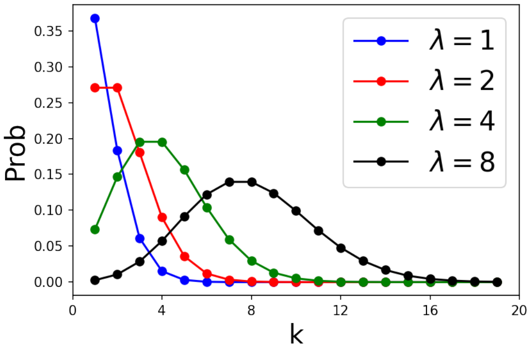

Figure 1: Degree distribution for average degree $\lambda = (n-1)p$.

We can use the Poisson approximation of the Binomial distribution to approximate the degree distribution of large graphs (i.e. $n \to \infty$):

$$P[deg(v) = k] \approx \frac{e^{-\lambda}\lambda^k}{k!}$$

where $\lambda = (n-1)p$ (expected degree).

Clustering coefficient:

$$C = p$$

For constant expected node degree $(n-1)p$, $C = (n-1)p/(n-1)$, which tends to 0 as $n \to \infty$

The diameter of an ER graph is $O(\log n)$. Let d_{uv} be the shortest path between nodes u and v. We know that the expected degree in G is $(n-1)p$ and thus the expected number of nodes within k hops from v is approximately $(n-1)^k p^k$ (for large enough n). Let $\ell = k_u + k_v - 1$, then $d_{uv} > \ell$ if the k_u-hop neighborhood of u is not connected to the k_v-hop neighborhood of v, thus:

$$P(d_{uv} > \ell) = (1-p)^{(n-1)^{\ell-1}p^{\ell-1}}$$

We are interested in the diameter as n grows and the average degree $(n-1)p = \lambda$ remains constant:

$$P(d_{uv} > \ell) = (1 - \frac{\lambda}{n-1})^{\lambda^{\ell-1}}$$

We can simplify the above equation taking the log of both sides and using the first term of the series expansion of $\log(1-\lambda/(n-1)) \approx -\lambda/(n-1)$:

$$P(d_{uv} > \ell) \approx \exp(-\frac{\lambda^\ell}{n-1})$$

By definition, the diameter of G is the minimum ℓ such that $P(d_{u,v} > \ell) = 0$. For large n, $P(d_{uv} > \ell) = 0$ when $\lambda^\ell = \Omega(n^{1+\epsilon})$ and $\epsilon > 0$. This requires $\ell = \mathit{const} + (1+\epsilon)\log(n)/\log(\lambda) = O(\log(n))$.

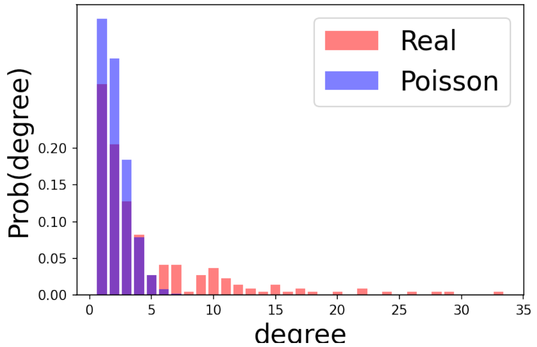

Figure 2: Degree distribution for a real co-mention network[4] and the predicted Poisson distribution.

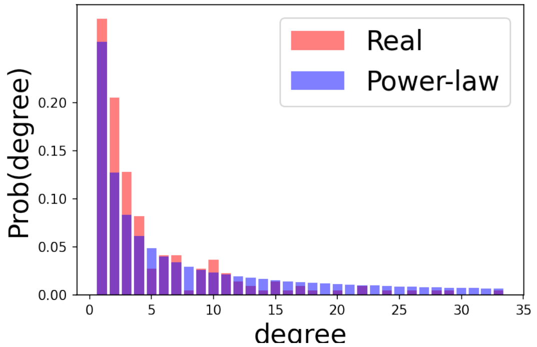

Figure 3: Degree distribution for a real co-mention network and the predicted power-law distribution.

Preferential Attachment

Preferential attachment models were motivated by some of the differences between ER and real networks. For instance, the degrees of the real networks described in the previous lecture do not fit a Poisson distribution. Figure 2 shows the degree distribution for a co-mention network among US congresspeople and the Poisson distribution with the best fit. Notice that the real distribution is much more *skewed*. A better fit for the degree distribution of our dataset with a *power-law distribution* is shown in Figure 3. The density function of the power-law distribution is defined as follows:

$$P(d_v = k) = \alpha k^{-\beta}$$

where α and β are parameters.

Preferential attachment models attempt to match the degree distributions observed in real data. Here, we will introduce the most popular preferential attachment model, the *Barabasi-Albert* (BA) network. The BA model shares many similarities with the *Price's model*, originally proposed for citation networks. The model starts with an empty network and nodes are added one by one. Each new node is connected to c existing ones selected with probability proportional to their current degree.

Degree distribution

Similar to the case of the ER networks, we will analyze the degree distribution of BA networks. Let the degree of node v be $d_v = c + q_v$, where q_v is the added degree of v—i.e. due to new nodes added to G. We know that the probability of such an added edge to be incident to v is:

$$\frac{c+q_v}{\sum_{u\in V} c+q_u} = \frac{c+q_u}{2nc}$$

Let $p_k(n)$ be the fraction of nodes with degree k when G has n nodes. The expected number of new edges connecting to nodes with added degree k is:

$$np_k(n) \times c \times \frac{k}{2nc} = p_k(n)\frac{k}{2}$$

We can use the above equation to write down a *master equation* for the evolution of the degree distribution as new nodes are added:

$$(n+1)p_k(n+1) = np_k(n) + p_{k-1}(n) \times \frac{k-1}{2} - p_k(n) \times \frac{k}{2}$$

which holds for $k > c$. For $k = c$, we have:

$$(n+1)p_c(n+1) = np_c(n) + 1 - p_c(n) \times \frac{c}{2}$$

Let p_k be the value of $p_k(n)$ as $n \to \infty$, then we can re-write the above equations as follows:

$$p_c = \frac{2}{2+c}$$

$$p_k = p_{k-1} \times \frac{k-1}{k+2}$$

Moreover, we can expand p_k and cancel several terms:

$$p_k = \frac{2}{2+c} \times \frac{k-q}{k-q+3} \ldots \times \frac{k-2}{k+1} \times \frac{k-1}{k+2} = 2 \times \frac{c(c+1)}{k(k+1)(k+2)}$$

Thus, for large k, we get:

$$p_k \approx k^{-3}$$

which shows that p_k follows a power-law with $\beta = 3$.

Other properties

The total number of edges in a BA network is $m = nc$ and the average degree is $2c$. It has been shown that the diameter of a BA network is approximately $\ln(n)/\ln\ln(n)$ and the clustering coefficient is approximately $\ln(n)^2/n$ [1].

References

[1] Albert-László Barabási et al. *Network Science*. Cambridge University Press, 2016.

[2] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[3] Mark Newman. *Networks*. Oxford university press, 2018.

[4] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, 2006.